Library and Information Science Research

https://Infosci.um.ac.ir Vol. 13, No 1, 2023 OPEN ACCESS

ISSN(Online): 2538-4112 ISSN(Print): 2251-6344 Research Article

Implementation of Experts' Retrieval Model Using Latent Semantic Indexing (LSA) Method and Temporal Graph

Shahla Rezvani

PhD graduate of Information Sciene and Knowledge Studies, University of Tehran, Tehran, Iran. Email: shahla.rezvani.gi@ut.ac.ir

Nader Naghshineh

Associate Professor of Information Science and Knowledge Studies, University of Tehran, Tehran, Iran. (Corresponding Author). Email: Nnaghsh@ut.ac.ir

Ahmad Khalilijafarabad

PhD in IT Management, Sahmeto Ltd., Tehran-Iran. Email: Ahmad.khalili@ut.ac.ir

Recived: 2020-02-03 Revised: 2023-01-03 Accepted: 2023-06-14 Published: 2023-06-21

Citation: Rezvani, S., Naghshineh, N., & Khalilijafarabad, A. (2023). Implementation of Experts' Retrieval Model Using Latent Semantic Indexing (LSA) Method and Temporal Graph. Library and Information Science Research, 13(1), 226-245. doi: 10.22067/infosci.2023.24058.0

Abstract

Introduction: Retrieval of experts is a subset of information retrieval that aims to provide a ranking of people who have knowledge in a particular field. Automated expertise work is challenging due to the abundance of expert information and data sources. Many expert approaches in both industry and academia have been proposed using new techniques in information retrieval, data mining, knowledge discovery, statistical modeling, probabilistic modeling, and complex networking. All researchers estimate the relationship between the query and the supporting documents of the expert candidate based on the occurrence of query words in the supporting documents, and they are main and important researches. These models are not capable of semantic communication. Therefore, in this research, the document-oriented method was considered using the LSA recovery model and the use of a time graph

Methodology: The research method is experimental ones, aside from this, survey and library methods have been used. The method used in current study to retrieve articles on LSA or Latent Semantic Analysis, which is based on the articles of the test collection prepared by Web of Science. These documents include English articles in information science and librarianship from 1989 to 2018 is indexed under the category of information science and librarianship on the website. Total number of these articles were 126924 and queries made by users were provided to all these articles. The retrieved documents were judged by relevance and after judging the relevance of the documents by the participants in the study, the performance of the information retrieval model was measured by the evaluation measurements of information retrieval systems. The result of the calculated measures was compared with the value of each of these measures in the basic model. A temporal graph was used to include the time factor. After that, the authors who had the most relevant work and their value of micro index of social network were introduced as experts. Then ten queries from the present research model and the basic model were randomly selected and given to eight people introduced by the second community for judgment and the results were



compared.

Findings: According to the innovation used in the current research, which was the application of the information retrieval model of latent semantic analysis, which was finally used to retrieve expert authors, in terms of the amount obtained from each of the information retrieval metrics, i.e., the accuracy level at the level of the first five results, or p@5, mean average precision (MAP) and mean inverse rank (MRR) with values of 0.895, 0.839 and 0.909, respectively, the latent semantic analysis recovery model performed better than the base model. In addition, this is due to the better performance of the retrieval using the dimensionality reduction method compared to keyword matching. In this method, hidden meaning indexing is used, which is a kind of conceptual indexing and uses the statistical method of least squares, and the above indexing is extracted by applying this statistical method. As we know, there are many ways to express a word (synonyms), so it is possible that the query words do not match the words of the document. In addition, most words have multiple meanings (multiple synonyms), so retrieving information based on the concept and meaning of a document is a better approach. LSI assumes that there is a number of latent structures in word usage that are partially blocked by diverse word choices. SVD is used to estimate this structure. The vectors that are obtained statistically strengthen the indicators of meaning more than individual words. The results of other researches also indicate that retrieving documents by matching query keywords with documents is a relatively weaker method. Also, the LSA retrieval model has a better performance in retrieving documents in a large set of documents than in a small set. According to the next innovation of the current research, which was the involvement of the time factor in expert search, and also according to the use of social network indicators and the final relevance judgment, the results showed that the performance of this method is significantly better than the model has been the base. The time factor was included in the retrieval of experts so that people who are no longer alive or who have been around for a long time since their last publication in a certain field are not retrieved. Considering the useful life of publications in the field of knowledge and information science, a ten-year period was involved. After using publication time as the determining factor of expert retrieval, those who had published the most related work were considered as the next determining factor and then the micro indicators of the social network such as degree centrality, betweenness centrality, closeness and special vector are other determining factors that are widely used in scientometric researches and recently in expert retrieval researches. The ten queries proposed in the current research were sent to 8 people who defined the second statistical population of the research, and the results indicated that the performance of the time graph and expert finding performed better by using the factor of the most relevant published works and the factor of micro-indexes of the social network.

Conclusion: LSI assumes that there is a number of latent structures in word usage that are partially blocked by diverse word choices. SVD is used to estimate this structure. The vectors that are obtained statistically strengthen the indicators of meaning more than individual words. The results of other researches also indicate that retrieving documents by matching query keywords with documents is a relatively weaker method. Also, the LSA retrieval model has a better performance in retrieving documents in a large set of documents than in a small set. According to the next innovation of the current research, which was the involvement of the time factor in expert search, and also according to the use of social network indicators and the final relevance judgment, the results showed that the performance of this method is significantly better than the model that has been the base. The time factor was included in the

retrieval of experts so that people who are no longer alive or who have been around for a long time since their last publication in a certain field are not retrieved. Considering the useful life of publications in the field of knowledge and information science and, a ten-year period was involved. After using publication time as the determining factor of expert retrieval, those who had published the most related work were considered as the next determining factor and then the micro indicators of the social network such as degree centrality, betweenness centrality, closeness and special vectors are other determining factors that are widely used in scientometric researches and recently in expert retrieval researches. is used. The results showed that the LSA model performed better than the base model for retrieving related documents and the use of time graph showed better performance than the base model.
Keywords: Latent Semantic Analysis, Temporal graph, Expert finding model, Time, Information system


پژوهش‌نامه کتابداری و اطلاع رسانی
مقاله پژوهشی https://infosci.um.ac.ir دسترسی آزاد

پیاده‌سازی مدل بازیابی خبرگان با استفاده از روش تحلیل معنای نهان و گراف زمان‌دار

شهلا رضوانی
دانش‌آموخته دکتری علم اطلاعات و دانش‌شناسی، دانشگاه تهران، تهران، ایران. shahla.rezvani.gi@ut.ac.ir
نادر نقشینه
دانشیار، علم اطلاعات و دانش‌شناسی، دانشگاه تهران، تهران، ایران. (نویسنده مسئول). Nnaghsh@ut.ac.ir
احمد خلیلی جعفرآباد
استادیار، مدیریت فناوری اطلاعات، دانشگاه تهران، تهران، ایران. Ahmad.khalili@ut.ac.ir

تاریخ دریافت: ۱۳۹۸/۱۱/۱۴ تاریخ بازنگری:۱۴۰۱/۱۰/۱۳ تاریخ پذیرش: ۱۴۰۲/۰۳/۲۴ تاریخ انتشار:۱۴۰۲/۰۳/۳۱

استناد: رضوانی, شهلا؛ نادر, نقشینه؛ خلیلی جعفرآباد, احمد. (۱۴۰۲). پیاده‌سازی مدل بازیابی خبرگان با استفاده از روش تحلیل معنای نهان و گراف زمان‌دار. پژوهشنامه کتابداری و اطلاع رسانی, ۱۳(۱), ۲۴۵-۲۲۶. doi: 10.22067/infosci.2023.24058.0

چکیده

مقدمه: خبره‌یابی شناسایی افراد با دانش و مهارت کافی در زمینه‌ای خاص و معرفی آنها به‌عنوان خبره در آن زمینه است. بازیابی افراد خبره زیرمجموعه‌ای از بازیابی اطلاعات است که هدف آن ارائه رتبه‌بندی از افرادی است که دارای دانش درزمینهٔ خاصی هستند. کار خبره‌یابی خودکار به‌دلیل فراوان‌بودن اطلاعات خبرگی و منابع داده چالش‌برانگیز است. هدف این پژوهش مقایسه عملکرد خبره‌یابی مدل بازیابی اطلاعات تحلیل معنای نهان و نیز گراف زمان‌دار با مدل پایه بود.

روش‌شناسی: روش پژوهش تجربی است و در کنار آن از روش کتابخانه‌ای نیز استفاده شده است. روشی که در پژوهش حاضر برای بازیابی مقالات استفاده می‌شود ال‌اس‌ای یا بازیابی معنای نهان است که بر روی مقالات مجموعه آزمون تهیه‌شده از وب‌آوساینس پیاده شد. این اسناد شامل مقالات انگلیسی علم اطلاعات و دانش‌شناسی است که از ۱۹۸۹ تا ۲۰۱۸ در پایگاه وب‌آوساینس در ذیل مقوله علم اطلاعات و دانش‌شناسی نمایه شده است. تعداد کل این مقالات ۱۲۶۹۲۴، پرس‌وجوهای ساخته‌شده توسط کاربران به همه این مقالات عرضه شد. اسناد بازیابی‌شده مورد قضاوت ربط قرار گرفتند و پس از انجام قضاوت ربط اسناد توسط شرکت‌کنندگان در پژوهش، عملکرد مدل بازیابی اطلاعات توسط سنجه‌های ارزیابی نظام‌های بازیابی اطلاعات اندازه‌گیری شد. سنجه‌های ارزیابی که در پژوهش حاضر مورداستفاده قرار گرفتند عبارت‌اند از میانگین متوسط دقت، میانگین معکوس رتبه، و دقت در سطح پنج نتیجه اول بازیابی شده. حاصل سنجه‌های محاسبه‌شده با مقدار هر یک از این سنجه‌ها در مدل پایه مقایسه شد. برای دخالت دادن عامل زمان از گراف زمان‌دار استفاده گردید. پس از دخالت دادن عامل زمان نویسندگانی که بیشترین کار مرتبط و نیز شاخص خرد شبکه اجتماعی آنها بیشتر بود به‌عنوان خبره معرفی گردید. سپس ده پرس‌وجو از مدل پژوهش حاضر و مدل پایه به‌طور تصادفی ساده انتخاب گردید و برای قضاوت در اختیار هشت نفر از افرادی که توسط جامعه دوم معرفی گردید قرار گرفت و نتایج حاصل باهم مقایسه گردید.

یافته‌ها: میزان به‌دست‌آمده از هر یک از سنجه‌های بازیابی اطلاعات یعنی میزان دقت در سطح پنج نتیجه اول، میانگین متوسط دقت (MAP) و میانگین معکوس رتبه (MRR) به‌ترتیب با مقدار ۰/۸۹۵، ۰/۸۳۹ و ۰/۹۰۹، مدل بازیابی تحلیل معنای نهان عملکرد بهتری نسبت به مدل پایه داشت؛ و این امر به‌دلیل بهتربودن عملکرد بازیابی به‌روش کاهش ابعاد نسبت

به تطابق کلیدواژه‌ای است. چون در این روش از نمایه‌سازی معنای نهان استفاده می‌شود که نوعی نمایه‌سازی مفهومی است و از روش آماری حداقل مربعات بهره می‌برد و نمایه‌سازی ذکرشده با به‌کارگیری این روش آماری استخراج می‌شود طبق تعریف پژوهشگران، خبره کسی است که بیشترین کار مرتبط با مجموع پرس‌وجوها در ده سال اخیر را داشته و دارای بالاترین مقدار در مرکزیت درجه‌ای، نزدیکی، بینابینی و بردار ویژه باشد. تعداد ۱۰ پرس‌وجو از هر پژوهش به‌طور مجموع ۲۰ پرس‌وجو به‌صورت اتفاقی انتخاب گردید و به خبرگان مشخص‌شده هر پژوهش توسط جامعه آماری سوم نمره صفر یا یک داده شد. مجموع نمرات برای هرکدام نشان می‌دهد دخالت‌دادن عامل زمان و استفاده از گراف زمان‌دار ازنظر نفر اول به میزان ۳ نمره و ازنظر نفر دوم نیز به‌اندازه ۳ نمره و ... از مدل پایه پیشی گرفته است.

نتیجه‌گیری: نتایج نشان دادند که مدل ال‌اس‌ای در مقایسه با مدل پایه جهت بازیابی اسناد مرتبط عملکرد بهتری داشته است و نیز استفاده از گراف زمان‌دار نسبت به مدل پایه عملکرد بهتری را نشان داده است.

کلیدواژه‌ها: تحلیل معنای نهان، گراف زمان‌دار، مدل بازیابی خبرگان، زمان، نظام اطلاعاتی

مقدمه

تفاوت مسئله بازیابی موجودیت با بازیابی اسناد در این است که موجودیت‌ها به‌طور مستقیم قابل‌نمایش مانند اسناد واحدهای قابل بازیابی نیستند، بنابراین باید آنها را به‌صورت غیرمستقیم و از طریق رخداد آنها در اسناد، شناسایی و بازیابی نمود که این مسئله منجر به ایجاد چالش‌های جدیدی در حوزه بازیابی اطلاعات می‌گردد (Balog, 2008). از میان انواع موجودیت‌های مختلف، افراد به‌عنوان یکی از پرارزش‌ترین منابع و موجودیت‌های وب محسوب می‌شوند. به همین علت، جستجوی افراد در سال‌های اخیر موردتوجه جامعه بازیابی اطلاعات قرار گرفته است. در واقع هدف، پاسخ‌گویی به این سؤال است؛ چه کسانی در موضوع x خبره هستند. پاسخ موردانتظار دسترسی به فهرست مرتب‌شده‌ای از افراد است که بر اساس میزان خبرگی خود در موضوع x رتبه‌بندی شده‌اند.

در حال حاضر بازیابی اطلاعات زمانی، حوزه پژوهشی درزمینهٔ بازیابی اطلاعات است. مفهوم استفاده از زمان در تعداد جستجو زیاد دارای اهمیت است (Mathews & Kanmani, 2012). پژوهش کان‌هابوا و نوواگ بر تحلیل گزارش‌های[1] پرس‌وجو، نشان می‌دهد که تعداد قابل‌توجهی از پرس‌وجوها زمانی هستند، منظور این است که وابسته به زمان هستند، و پرس‌وجوهای زمانی در بسیاری از حوزه‌ها همانند کتابخانه‌های دیجیتال و آرشیوهای اسناد نقش مهمی را بازی می‌کنند. پرس‌وجوهای زمانی به دو دسته تقسیم می‌شوند: ۱. آنهایی که معیار زمانی معینی را ارائه می‌کنند؛ و ۲. آنهایی که بدون معیار زمانی ارائه می‌شوند (Kanhabua & Nørvåg, 2010). منظور پژوهش حاضر است که کاندیداهای بازنشسته یا فوت‌شده بازیابی نشوند؛ چراکه کاندیداهای بازنشسته‌شده ممکن است دیگر در حیطه موردنظر انتشاراتی

1. logs

نداشته باشند و درنتیجه اطلاعات روزآمدی در اختیار کاربران قرار نگیرد. با توجه به تغییرات سریعی که در دنیای اطلاعات رخ می‌دهد و گذشت چند سال از عمر مفید اطلاعات می‌کاهد لازم است تا کاندیداهایی بازیابی شوند که دارای انتشارات روزآمد در موضوع موردنیاز کاربران باشند؛ این امر در مورد کاندیداهایی که در قید حیات نیستند نیز صادق است. در این پژوهش بازیابی خبرگان با استفاده از مدل بازیابی تحلیل معنای نهان یا ال‌اس‌ای انجام گرفت (بدین منظور مجموعه آزمونی توسط محققان ساخته شد) و عملکرد آن را با استفاده از سنجه‌های بازیابی به دست آورده و با عملکرد مدل بازیابی DLH13 که یکی از چند مدل بازیابی استفاده‌شده در رساله هاشم عطاپور بود و به‌عنوان مدل بازیابی که عملکرد بهتری داشت معرفی شده بود مقایسه شد؛ و درنهایت از گراف زمان‌دار (تا کاندیداهای ده سال اخیر معرفی شوند) همراه با سنجه‌های مرکزیت برای تعیین خبرگان نهایی استفاده شد و نتیجه دوباره با مدل پایه مورد مقایسه قرار گرفت.

مبانی نظری

تحلیل معنای نهان

روشی برای بازیابی اطلاعات است که برای غلبه بر مشکلات بازیابی اطلاعات طراحی شده است. روش LSA گاهی در بازیابی اطلاعات با LSI[1] یعنی نمایه‌سازی معنای نهان شناخته می‌شود. نمایه‌سازی معنای نهان یکی از راهکارهایی بود که برای رفع مشکلات مدل فضابرداری به وجود آمد. مدل فضابرداری یکی از چندین روش تشخیص تشابه میان دو مدرک است؛ که در سال ۱۹۷۵ توسط سالتون گسترش یافت (Salton et al., 1975). مدل فضابرداری با عبارت‌های نامشابه مانند اقلام نامرتبط برخورد می‌کند. این موضوع یعنی نبود مطابقت واژگان اصلی‌ترین مشکل مدل فضابرداری است. دومین اشکال این مدل تشکیل ماتریس عبارت_ مدرک بزرگ و پراکنده‌ای است که در مجموعه بزرگی از مدارک که فضای ذخیره زیادی را می‌طلبد و مدت‌زمان پردازش و محاسبه آن نیز طولانی است وجود دارد. برای رفع این مشکلات از یکی از شاخه‌های این مدل با عنوان نمایه‌سازی معنای نهان استفاده می‌شود که در سال‌های اخیر رواج یافته است (Chaomei & Paul, 2001; Cheng, 2003). نمایه‌سازی نهان اولین بار توسط گروهی از پژوهشگران به‌نام دیروستر و همکارانش در شرکت بل[۲] در بازیابی اطلاعات به کار رفت (Deerwester et al., 1990) و نمایه‌سازی نهان نامیده شد. دلیل پژوهشگران از به‌کارگیری نمایه‌سازی نهان آن است که عبارت‌های جدید که بازنمون اطلاعات معنایی هستند به‌صورت مستقیم از مدارک یافت نمی‌شود، بلکه حاصل بررسی مجموعه مدارک و استفاده از روش ریاضی با عنوان SVD[3] یا تجزیه مقادیر منفرد است.

1. latent semantic indexing
2. Bellcore
3. singular value decomposition

گراف زمان‌دار

گراف زمان‌دار، ساختار داده‌ای شامل گره‌ها و یال‌هاست که در آن یال‌ها با برچسب زمان ارتباط دارند (Lightenber & Pie, 2017) (Ligtenberg & Pei, 2017). گراف زمان‌دار به‌صورت غیررسمی نموداری است که با زمان تغییر می‌کند وقتی زمان گسسته است و تنها روابط بین موجودیت‌های شرکت-کننده ممکن است تغییر کند و نه خود موجودیت‌ها (Michail, 2015) (Michail, 2015). هم‌اکنون مجموعه‌ای عظیم از کاربردها و حوزه‌های پژوهشی شناسایی شده است که از انگیزه و توانایی لازم جهت توسعه مجموعه واقعی از نتایج، ابزارها و شیوه‌ها برای استفاده از گراف زمان‌دار برخوردارند. انواع زیادی از هر دو شبکه مدرن‌وسنتی مانند شبکه‌های اطلاعاتی و ارتباطی، شبکه‌های اجتماعی، شبکه‌های حمل‌ونقل، و چندین نظام فیزیکی می‌توانند به‌طور طبیعی به‌عنوان گراف زمان‌دار مدل‌سازی شوند.

پیشینه پژوهش

رویکردهای کلان اولیه برای بازیابی تخصص در قالب پایگاه‌هایی ظاهر شد که حاوی بازنمون‌های دانش و مهارت نیروی کار سازمان بود. ایجاد پرونده کارکنان در نظام‌های اولیه دستی، و به‌طور معمول تکمیل و روزآمدسازی آن برعهدهٔ خود کارکنان بود. این امر، بارکاری زیادی را بر کارکنان و مدیران نظام‌ها تحمیل می‌کرد. به‌دلیل مشکلاتی که نظام‌های دستی بازیابی تخصص داشتند، تلاش‌ها در جهت توسعه نظام‌های خودکار بازیابی تخصص افزایش یافت. اولین تلاش‌ها برای خبره‌یابی خودکار را می‌توان در پژوهش‌هایی ردیابی کرد که سعی می‌کردند خبرگان نهفته در ارتباطات با رایانامه را کشف کنند. ازآنجاکه اسناد ایمیلی به‌طور طبیعی فعالیت‌ها و علایق افراد را در برداشتند، برای عمل خبره‌یابی مناسب قلمداد می‌شدند. پژوهش شوارتز و وود[1] (۱۹۹۳) اولین پژوهشی است که برای یافتن افراد، به مشاهده الگوهای ارتباطات ایمیلی پرداخته است (Schwartz & Wood, 1993). در آن پژوهش، فیلدهای «از» و «به» از لاگ‌های رایانامه ۱۵ وب‌سایت به‌مدت دو ماه جمع‌آوری شد و برای تولید گرافی که در حدود ۵۰ هزار نفر را در برداشت، مورداستفاده قرار گرفت. سپس تعدادی از الگوریتم‌های اکتشافی گراف برای خوشه‌بندی افراد دارای علایق مشترک به کار گرفته شد: یعنی زیرمجموعه‌ای از گره‌ها که ارتباطات بسیار بالایی باهم داشتند، مورد جستجو قرار گرفت.

کمپل، مگیلو، کوزی و دان[2] (۲۰۰۳) و عسکری، وربرنی و پاسی[3] (۲۰۲۲) هم در پژوهش خود رایانامه را از سازمان پژوهشی و سازمان توسعه نرم‌افزاری گردآوری کردند (Campbell et al., 2003)، (Askari et al., 2022). الگوریتم گراف‌محور هیتس با الگوریتم محتوا‌محور مقایسه شد. فو، زیانگ، ژانگ،

1. Schwartz & Wood
2. Campell, Magilo, Cozzi, & Down
2. Askari, Verberne, & Pasi

لیو و ما[1] (۲۰۰۶) رویکرد جستجوی توصیف اسناد شخص‌ـمحور[2] را برای خبره‌یابی پیشنهاد کردند (Fu et al., 2006). این مدل نشان داد که مؤثر بوده و به بهترین عملکرد در مجموعه آزمون ترک[3] ۲۰۰۵ دست یافته است. فنگ و زای[4] (۲۰۰۷) به ارائه مدل‌های احتمالاتی تشخیصی برای منابع اطلاعات ناهمگن جهت خبره‌یابی پرداختند؛ و چهارچوب یادگیری، با ناظر ربط پایه برای یادگیری ترکیب وزن‌ها از داده‌های آموزشی ارائه شد. نتایج نشان‌دهنده کارآمدی و توان مدل‌های یادگیری تشخیصی بودند (Fang & Zhai, 2007). ژانگ، تانگ و لی[5] (۲۰۰۷) در پژوهشی شبکه دانشگاهی که در آن هر شخص می‌تواند انواع متفاوتی از اطلاعات از قبیل پرونده شخصی، اطلاعات تماس، و انتشارات/اسناد و روابط در شبکه هم تألیفی داشته باشد را به‌عنوان داده تجربی انتخاب کردند (Jing Zhang et al., 2007). راهبرد آنها برای محاسبه امتیازات اولیه خبرگی بر اساس مدل احتمالاتی بازیابی اطلاعات بود. ژو، اورشانسکی، ژای و گایلز [6] (۲۰۰۷) در پژوهشی به مسئله رتبه‌بندی اسناد و نویسندگان پرداختند (Zhou et al., 2007). آنها الگوریتم خود را بر روی مقالات گردآوری‌شده از پایگاه CiteSeer آزمون کردند. قضاوت ربط توسط خود پژوهشگران در کنار فرد دیگری انجام شد. سردیوکوف و هیمسترا (۲۰۰۸) در پژوهش خود این‌گونه اذعان داشتند که وقوع اصطلاحات در اسناد مستقل از حضور کاندیدا خبره نمی‌تواند باشد. کاندیداها با استفاده از مدل زبانی ساخته‌شده از اسناد بازیابی‌شده و بسامد حضور یک کاندیدا در اسناد بازیابی‌شده در رتبه بالایی قرار می‌گیرد. در این پژوهش آزمایش‌های بر روی مجموع آزمون ترک انجام شد و نشان داده شد که این رویکرد در مقایسه با بهترین راه‌حل‌های موجود برتری دارد (Serdyukov & Hiemstra, 2008). بالوگ[7] (۲۰۰۸) در پایان‌نامه خود آزمایش‌های مربوط به خبره‌یابی را در مجموعه‌های ترِک ۲۰۰۷-۲۰۰۵ و آزمایش‌های مربوط به تشکیل پروفایل خبرگان را در مجموعه آزمون دانشگاه تیلبورگ پیاده‌سازی کردند. نتایج این آزمایش‌ها نشان داد که مدل سندـمحور هم در خبره‌یابی و هم در تشکیل پروفایل خبرگان عملکرد بهتری از مدل کاندیداـمحور دارد (Balog, 2008). بوگرس، کوکس و وان دن بوش[8] (۲۰۰۸) در پژوهش خود تأثیر ترکیب دو منبع مختلف خبرگی یعنی محتواـپایه و شبکه اجتماعی را برای یافتن خبره در کارگروه دانشگاهی متوسط بررسی کردند (Bogers et al., 2008). مجموعه آزمون مورداستفاده در این پژوهش ILK[9] بود. این مجموعه در حقیقت زیرمجموعه‌ای از

1. Fu, Xiang, Zhang, Liu, & Ma
2. person description document-based
3. trec
4. Fang & Zai
5. Zhang, Tang, & Li
6. Zhou, Orshanskiy, Zha, & Giles
7. Balog
8. Bogers, Kox & Van Den Bosch
9. induction of linguistic knowledge test collection

مجموعه بازیابی خبرگان UVT[1] است. این پژوهشگران رویکرد سند-محور بازیابی خبرگان، که نوعی رویکرد محتوا محور است را به‌عنوان سطح پایه اتخاذ کردند. سپس به بررسی مرکزیت درجه درونی و پیج رنک[۲] نویسندگان در شبکه استنادی مقالات و همچنین شبکه استنادی نویسندگان پرداختند. ژانگ، تانگ، لیو و لی[۳] (۲۰۰۸) مدل ترکیبی برای بازیابی خبره با استفاده از موضوع‌ها ارائه دادند که در آن اطلاعات معنایی نیز وارد فرایند بازیابی خبرگان می‌گردد (Zhang et al., 2008). مدل ترکیبی، بازنمایی از لایه پنهان بین کلمات و اسناد پشتیبان با استفاده از تحلیل معنای نهان احتمالاتی یا PLSA است. مک‌دونالد [۴] (۲۰۰۹) در پایان‌نامه دکتری خود از مدل رأی‌دهی استفاده کرد (Macdonald, 2009). وقتی سندی برای پرس‌وجو بازیابی می‌شود این سند رأیی از هر شخصی که با آن ارتباط دارد است. در واقع برای هر فردی که با سند در ارتباط است یک رأی محسوب می‌شود. پارادایم رأی‌دهی، به چگونگی تلفیق آرائی می‌پردازد که اسناد به کاندیداها اعطا کرده‌اند. دائود، لی، ژو و موهمد[5] (۲۰۱۰) در پژوهش خود مسئله بازیابی خبرگان زمانی معنا پایه را موردتوجه قرار دادند که به‌معنی شناسایی شخصی با تخصص خاص در دوره‌های زمانی مختلف است (Daud et al., 2010). درحالی‌که معنی موضوعی خاص کم‌وبیش بدون تغییر می‌ماند. نتایج به‌طور معناداری نشان‌دهنده برتری رویکرد مدل‌سازی کلی زمان موضوع بر رویکرد مدل‌سازی که مبتنی بر فقدان تأثیر کنفرانس و زمان اطلاعات بود است. امیدوار، گراکانی و صفرپور[۶] (۲۰۱۴)، به‌دلیل اینکه مشکل اصلی در مورد اشتراک دانش در فوروم‌های مقدار زیاد داده بدون هیچ سازوکاری برای تعیین اعتبار آنها است، بنابراین برای جستجوگران، دانش سطح خبرگی هر عضو در یک زمینه خاص مهم است. درنهایت به‌منظور یافتن پاسخ‌های معتبر الگوریتمی به‌نام رتبه‌بندی محتوا محور کاربر پیشنهاد شد (Omidvar et al., 2014).

عطاپور (۱۳۹۴)، در رساله خود برای یافتن نویسندگان خبره مدل سند-محور خبره‌یابی را مبنای کار قرار داد و در بطن آن از مدل‌های رأی‌دهی و رویکرد گراف‌محور نیز استفاده کرد. این مدل شامل دو مرحله بازیابی سند و تخصیص اسناد به افراد است. بدین منظور مجموعه آزمونی متشکل از ۵۵ پرس‌وجو و ۹۶۳۷۵ مقاله نمایه‌شده علم اطلاعات و دانش‌شناسی در پایگاه وب‌آوساینس ایجاد شد. پرس‌وجوها به پیکره آزمون عرضه شد و ۶ مدل مختلف بازیابی اطلاعات برای بازیابی اسناد به کار گرفته شد. اسناد بازیابی‌شده در معرض قضاوت ربط قرار گرفته، سپس عملکرد مدل‌های بازیابی اطلاعات موردسنجش و مقایسه قرار گرفتند. پس از مقایسه عملکرد مدل‌های بازیابی اطلاعات، نتایج مدل DLH13 برای مرحله

1. tilburg university expert collection
2. page rank
3. Zhang, Tang, Liu & Li
4. Macdonald
5. Daud, Li, Zhou, & Muhammad
6. Omidvar, Garakani, & Safarpour

دوم انتخاب شد. اسامی افراد موجود در اسناد بازیابی‌شده توسط مدل منتخب استخراج و پردازش شد. سپس با به‌کارگیری ۱۶ روش مختلف رتبه‌بندی گردید. ۱۰ نتیجهٔ اول هر روش انتخاب و در انبوه نویسندگان پرس‌وجوی مربوطه قرار گرفت. افراد موجود در سیاهه مورد قضاوت ربط قرار گرفتند تا مقایسه عملکرد مدل‌های مختلف خبره‌یابی میسر شود.

نتیجه پژوهش کمپل و همکارانش نشان داد، استفاده از امتیازات نفوذ الگوریتم هیتس برای رتبه‌بندی کاندیداها، در مقایسه با روش محتوا-محور از دقت بالاتر و جامعیت کمتری برخوردار است. علاوه بر ارتباطات رایانامه، فنون تحلیل شبکه‌های اجتماعی برای امر خبره‌یابی در سایر منابع ایجاد شبکه‌های اجتماعی نیز به کار گرفته شده است. این منابع شامل گزارش‌های چت (Ehrlich et al., 2007)، فوروم‌های بحث (Jun Zhang et al., 2007)، و اطلاعات شبکه‌های هم‌تألیفی در پایگاه‌های کتابشناختی (Li et al., 2007; Jing Zhang et al., 2007) بود. در پژوهش بوگرس، کوکس و وان دن بوش (۲۰۰۸) یکی از نتایج پژوهش این بود که عملکرد رتبه‌بندی ایستا (مستقل از پرس‌وجو) تفاوت معناداری با رتبه‌بندی وابسته به پرس‌وجو ندارد (Bogers et al., 2008). توجیه پژوهشگران این است که چون مجموعه آزمون مورداستفاده بسیار تخصصی است، همان نویسندگانی که در رابطه با موضوع پرس‌وجو رتبه‌های خوبی به دست می‌آورند، به‌لحاظ شاخص‌های مرکزیت درونی و پیج‌رنک در شبکه استنادی نیز رتبه‌های خوبی به دست می‌آورند. یکی دیگر از نتایج پژوه آنها این بود که استفاده از شبکه استنادی، در عملکرد خبره‌یابی بهبود معناداری ایجاد نمی‌کند.

مدل خبره‌یابی را می‌توان در چهار گروه کلی مدل‌های احتمالاتی زایشی، مدل احتمالاتی تشخیصی، مدل‌های مبتنی بر گراف، و مدل‌های رأی‌گیری تقسیم‌بندی کرد. جستجوی خبره بسیار نزدیک و مرتبط به جستجوی سند و دیگر وظایف بازیابی اطلاعات است؛ بنابراین برخی از روش‌های کلاسیک بازیابی اطلاعات در فرایند بازیابی خبره نیز استفاده می‌گردد که می‌توانند به‌عنوان مدلی دیگر در فرایند بازیابی خبره دیده شوند. یکی از این روش‌ها، مدل کردن موضوع‌ها و استفاده از آنها به‌عنوان لایه‌ای پنهان در فرایند بازیابی است.

تمامی پژوهش‌های موردبررسی در این بخش (به‌جز پژوهش ژانگ و همکاران[1] (۲۰۰۸)) ارتباط بین پرس‌وجو و اسناد پشتیبان کاندیدای خبرگی را بر اساس رخداد کلمه‌های پرس‌وجو در اسناد پشتیبان تخمین می‌زنند و از پژوهش‌های اصلی و مهم در بازیابی خبرگان هستند. این مدل‌ها قادر به ارتباط معنایی نیستند. بنابراین در پژوهش حاضر روش سند-محور با استفاده از مدل بازیابی ال‌اس‌ای و به‌کارگیری گراف زمان‌دار (با استفاده از زبان برنامه‌نویسی پایتون[2]) در نظر گرفته شد.

1. Zhang et al.
2. python

اهداف اصلی پژوهش

هدف اصلی پژوهش حاضر پیاده‌سازی مدل بازیابی خبرگان با استفاده از روش ال‌اس‌ای و گراف زمان‌دار است.

اهداف فرعی پژوهش

۱. شناسایی روش بازیابی اطلاعات تحلیل معنای نهان یا ال‌اس‌ای به‌منظور بازیابی اسناد؛
۲. شناسایی کاربرد گراف زمان‌دار و دخالت دادن عامل زمان در رتبه‌بندی خبرگان بازیابی‌شده؛

سؤالات پژوهش

۱. آیا مدل بازیابی ال‌اس‌ای از عملکرد مناسب‌تری نسبت به مدل پایه برخوردار است؟
۲. آیا دخالت دادن عامل زمان و استفاده از گراف زمان‌دار نسبت به مدل پایه از عملکرد بهتری برخوردار است؟

روش‌شناسی پژوهش

جامعه و نمونه پژوهش

پژوهش حاضر دارای سه جامعه است: الف) اسناد مجموعه آزمون؛ ب) افرادی که به ساخت پرس‌وجوها و قضاوت ربط در مورد اسناد بازیابی‌شده می‌پردازند؛ و ج) افرادی که به قضاوت ربط در مورد خبرگان بازیابی‌شده می‌پردازند. مجموعه آزمون مشتمل بر مجموعه بزرگی از اسناد بوده و دارای سه مؤلفه است: مجموعه اسناد، مجموعه‌ای از عناصر موضوعی (پرس‌وجوها)، و مجموعه‌ای از قضاوت ربط (Sanderson, 2010). مجموعه اسناد که بدنه اصلی مجموعه آزمون را تشکیل می‌دهند، جامعه اول پژوهش هستند. این اسناد شامل مقالات انگلیسی علم اطلاعات و دانش‌شناسی است که از ۱۹۸۹ تا ۲۰۱۸ در پایگاه وب‌آوساینس در ذیل مقوله علم اطلاعات و دانش‌شناسی نمایه شده است. تعداد کل این مقالات ۱۲۶۹۲۴ بود و پرس-وجوهای ساخته‌شده توسط کاربران به همه این مقالات عرضه شد و در بین مقالات ذکرشده عمل نمونه-گیری صورت نگرفت. برای ساختن پرس‌وجوها و انجام قضاوت‌های ربط در مورد اسناد و نویسندگان بازیابی‌شده، باید افرادی انتخاب شوند که از حوزه موضوعی پرس‌وجو و نویسندگان خبره شناخت داشته و توانایی انجام قضاوت ربط در مورد اسناد بازیابی‌شده را داشته باشند. به این دلیل فارغ‌التحصیلان و دانشجویان تحصیلات تکمیلی رشته علم اطلاعات و دانش‌شناسی (گرایش بازیابی اطلاعات) دانشگاه تهران از سال ۱۳۹۴ تا ۱۳۹۷ به‌عنوان یکی از سه جامعه پژوهش انتخاب شدند. درمجموع ۱۵۰ پرس‌وجو از ۵۳ نفر گردآوری شد. جامعه سوم پژوهش، افرادی هستند که به قضاوت ربط در مورد نویسندگان خبره بازیابی‌شده برای هر پرس‌وجو می‌پردازند. نیاز هست که این افراد از موضوع پرس‌وجو و نیز افراد خبره در آن موضوع شناخت جامعی داشته باشند. در این پژوهش برای قضاوت ربط در مورد نویسندگان خبره بازیابی‌شده، از هشت نفر که توسط افراد جامعه دوم معرفی می‌شوند استفاده خواهد شد.

روش پژوهش

پژوهش حاضر پژوهشی کاربردی است. پژوهش حاضر پژوهشی کاربردی و کیفی است. در تقسیم‌بندی نوعی، پژوهش تجربی به سه نوع حقیقی، مقدماتی و نیمه تجربی تقسیم شده است. پژوهش حاضر نیز با ویژگی‌های مطالعه موردی تک‌مرحله‌ای از زیرمجموعه‌های پژوهشی تجربی مقدماتی همخوانی دارد و می‌توان آن را پژوهش تجربی مقدماتی در نظر گرفت.

مدل پایه‌ای که یافته‌های پژوهش حاضر با یافته‌های آن مورد مقایسه قرار می‌گیرد مدل برتر به‌دست‌آمده در رساله دکتری هاشم عطاپور است. پژوهش وی شامل سه مرحله تشکیل مجموعه آزمون بازیابی اسناد، استخراج اسامی و تشکیل پرونده نویسندگان، و قضاوت ربط در مورد خبرگی نویسندگان است؛ که پیش‌تر در بخش پیشینه ذکرشده است. مراحل پژوهش حاضر در بخش فرایند اجرای پژوهش ذکر خواهد شد اما تفاوت اصلی در روش پژوهش آنها در استفاده از روش بازیابی اسناد تحلیل معنای نهان[1] و گراف زمان‌دار[2] است که در مجموعه آزمون عامل زمان در رتبه‌بندی کاندیداهای بازیابی دخالت داده شده است.

دقت در سطح پنج نتیجه اول دقت، عبارت است از نسبت اسناد مربوط بازیابی‌شده به‌کل اسناد بازیابی‌شده برای یک پرس‌وجو (Baeza-Yates & Ribeiro-Neto, 2012). به‌طورمعمول، دقت در بین n سند اول بازیابی‌شده محاسبه می‌شود. مقادیر n به‌طور معمول عبارت‌اند از ۵، ۱۰، ۱۵، ۲۰، ۳۰، و ۱۰۰. دقت در سطح نتایج اول بسیار مهمتر از دقت در نتایج پایین‌تر محسوب می‌شود، زیرا بازیابی نتایج نامرتبط در نتایج اول به از دست رفتن زمان کاربر شده و چه‌بسا هزینه‌هایی بر وی تحمیل می‌کند. در پژوهش حاضر، دقت در سطح ۵ نتیجه اول (n=5) مدنظر قرار می‌گیرد. برای محاسبه دقت در سطح ۵ نتیجه اول، کافی است تعداد آیتم‌های مرتبط در ۵ نتیجه اول را بر عدد ۵ تقسیم کرد. میانگین متوسط دقت[3]: عبارت است از مجموع امتیازات متوسط دقت برای هر پرس‌وجو تقسیم بر تعداد کل پرس‌وجوها. میانگین معکوس رتبه[4]: منظور از معکوس رتبه پرس‌وجو، معکوس رتبه اولین پاسخ صحیح (مربوط) است. میانگین معکوس رتبه، از تقسیم مجموع معکوس رتبه همه پرس‌وجوها بر تعداد کل پرس‌وجوها به دست می‌آید.

فرایند اجرای پژوهش

به‌دلیل دخالت دادن عامل زمان، مجموعه مقالات علم اطلاعات و دانش‌شناسی که از سال ۱۹۸۹ به بعد در پایگاه وب‌آوساینس در ذیل مقوله علم اطلاعات و دانش‌شناسی نمایه شده‌اند، برای بررسی انتخاب شدند. مزیت انتخاب مقالات این حوزه به‌عنوان مجموعه مدارک آزمون، آشنایی پژوهشگران با

1. latent semantic analysis
2. temporal graph
3. mean average precision (MAP)
4. mean reciprocal rank (MRR)

مفاهیم آن است که می‌تواند در تحلیل آزمایش‌های نهایی کمک‌کننده باشد. ازآنجاکه جامعه آماری ایجادکننده مجموعه پرس‌وجو اغلب به زبان انگلیسی آشنایی داشته و پرس‌وجوهای خود را با زبان انگلیسی مطرح می‌کنند، بنابراین فقط باید به مقالات انگلیسی اکتفا کرد. برای دریافت مقالات مذکور، با عنایت به مقوله‌بندی‌های وب‌آوساینس در بخش جستجوی پیشرفته این پایگاه عبارت: =Wc Information science & Library science تایپ شد و سال‌های انتشار مقالات به سال ۱۹۸۹ تا ۲۰۱۸ محدود گردید. Wc نشانگر Web of science category است.

تعداد کل این مقالات ۱۲۶۹۲۴ بود. عملیات پردازش متن، ریشه‌یابی و حذف کلمات زائد بر روی پیکره اسناد اعمال شد؛ و از افراد خواسته شد فرم پرس و جوی خود را طی ایمیلی که به آنها ارسال می‌شود کامل کنند و پرس‌وجوی خود را در حوزه‌هایی که در آن شناخت دارند مطرح کنند. یک نمونه از پرس‌وجو در جدول (۱) آمده است.

جدول ۱. نمونه‌ای از پرس‌وجو

ردیف	عنوان پرس‌وجو	توصیف	شرح پرس‌وجو	خبره یا خبرگان حوزه
مثال	Cybernetics and human life-style	Information on the effects of cybernetics on human life	Information on the various uses of cybernetics in human life and how they could change the human lifestyle	Norbert Wiener Ashby Strawberry Bear McCulloch

در دومین گام، از مدل بازیابی تحلیل معنای نهان یا ال‌اس‌ای برای بازیابی مقالات مربوط به پرس-وجوی جامعه استفاده شد؛ که از شاخص‌های سنجش کیفیت (مربوط و نامربوط) جهت انجام قضاوت ربط توسط جامعه آماری استفاده گردید؛ و از نرم‌افزار اکسل برای محاسبه و شمارش سنجه‌های ارزیابی نظام بازیابی اطلاعات (میانگین متوسط دقت، میانگین معکوس رتبه، و دقت در سطح پنج رتبه اول) استفاده شد.

گام سوم مقایسه نتایج بازیابی‌شده توسط ال‌اس‌ای با نتایج بازیابی‌شده مدل پایه است. گام چهارم طبق تعریف نویسندگان از خبره، که خبره کسی است که دارای بیشترین کار مرتبط در ارتباط با مجموع پرس‌وجوها داشته دارای بالاترین مقدار در مرکزیت درجه‌ای، نزدیکی، بینابینی و بردار ویژه باشد و نیز مربوط به ده سال اخیر باشد را که همان دخالت دادن گراف زمان‌دار است برای رتبه‌بندی استفاده می‌شود. در مراحل برنامه‌نویسی از نرم‌افزار پایتون استفاده می‌شود.

در گام آخر به‌صورت اتفاقی تعدادی نتیجه پرس‌وجو به جامعه سوم نشان داده می‌شود که ترکیب خروجی پژوهش حاضر و مدل هاشم عطاپور باشد و از افراد بخواهیم باهم مقایسه کنند و سپس بررسی خواهد شد که کار حاضر چه اندازه بهتر بوده است.

پایایی و روایی پژوهش

محاسبه دستی چند مورد از نویسندگان و مقایسه آنها با امتیاز محاسبه‌شده توسط نرم‌افزار صحت کدهای نوشته‌شده را تأیید می‌کند. قابلیت اعتماد یا پایایی یکی از ویژگی‌های اندازه‌گیری است و با این امر سروکار دارد که ابزار اندازه‌گیری در شرایط یکسان تا چه اندازه نتایج یکسانی به دست می‌دهد. برای محاسبه ضریب اعتماد ابزار اندازه‌گیری میزان ربط شیوه‌های مختلفی به کار برده می‌شود؛ از آن جمله می‌توان به اجرای دوباره (روش بازآزمایی)، روش موازی، روش تنصیف (دونیمه کردن) و روش کودریچاردسون و آلفای کرانباخ اشاره کرد (Sarmad et al., 2023). برای سنجش قضاوت‌های ربط اسناد، از روشی استفاده شده که به‌روش آزمون‌های همتا[1] شبیه است. بدین ترتیب که در خلال نتایج بازیابی‌شده، نتایج تکراری جا گرفت تا میزان دقت و پایایی قضاوت‌کننده مشخص گردد. میزان همبستگی به‌دست‌آمده در این روش بسیار بالا (۰/۹۸) و حاکی از پایایی نتایج است. برای سنجش قضاوت ربط نویسندگان خبره روش آزمون‌های موازی (قضاوت هم‌زمان چند نفر) به کار گرفته شد. میزان همبستگی به‌دست‌آمده در امر قضاوت ربط نویسندگان ۰/۸۳ به دست آمد که حاکی از پایا بودن قضاوت‌های ربط انجام شده است.

یافته‌ها

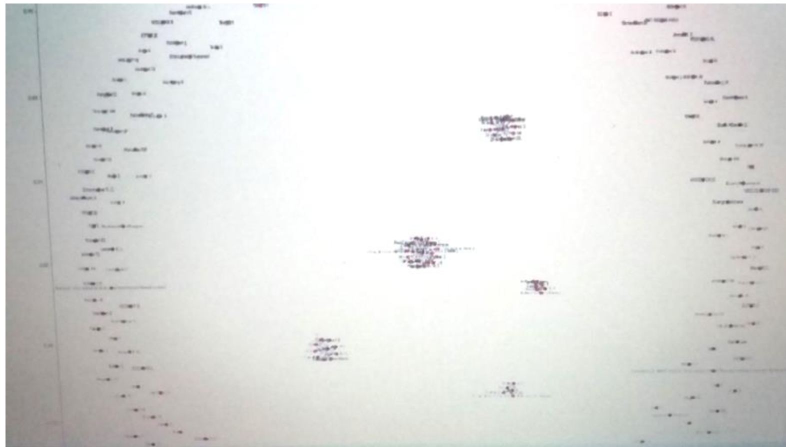
آیا مدل بازیابی LSA از عملکرد مناسب‌تری نسبت به مدل پایه برخوردار است؟ برای پاسخ به این سؤال پژوهش از سه سنجه میانگین متوسط دقت، میانگین معکوس رتبه، و دقت در سطح پنج رتبه اول استفاده شد. در جدول (۲) نتیجه به‌دست‌آمده از محاسبه این سه سنجه در مدل ال‌اس‌ای[2] با مدل پایه که DLH13 است مقایسه می‌گردد؛ که به‌ترتیب دقت در سطح پنج نتیجه اول در مدل ال‌اس‌ای از مدل پایه به‌اندازه ۰/۶۰ عملکرد بهتری داشته است؛ و میانگین متوسط دقت در مدل ال‌اس‌ای به‌اندازه ۰/۲۶۴ عملکرد بهتری را نشان می‌دهد؛ و بالاخره میانگین معکوس رتبه در مدل ال‌اس‌ای به‌اندازه ۰/۱۰۰ عملکرد بهتری را داشته است.

جدول ۲. مقایسه سه سنجه در مدل ال‌اس‌ای و مدل پایه

دقت در سطح پنج نتیجه اول در مدل LSA	میانگین متوسط دقت در مدل LSA	میانگین معکوس رتبه در مدل LSA	دقت در سطح پنج نتیجه اول در مدل DLH13	میانگین متوسط دقت در مدل DLH13	میانگین معکوس رتبه در مدل DLH13
۰/۸۹۳	۰/۸۳۱	۰/۹۰۴	۰/۸۸۷	۰/۵۶۷	۰/۹۰۳

1. matching method
2. LSA (Latent semantic analysis)

آیا دخالت دادن عامل زمان و استفاده از گراف زمان‌دار نسبت به مدل پایه از عملکرد بهتری برخوردار است؟ طبق تعریف پژوهشگران، خبره کسی است که بیشترین کار مرتبط با مجموع پرس‌وجوها در ده سال اخیر را داشته و دارای بالاترین مقدار در مرکزیت درجه‌ای، نزدیکی، بینابینی و بردار ویژه باشد. همان گونه که در شکل شماره (۱) مشاهده می‌شود نویسندگانی که سال انتشار آثارشان در بازه زمانی ده سال نبود در خارج از گراف قرار گرفتند.

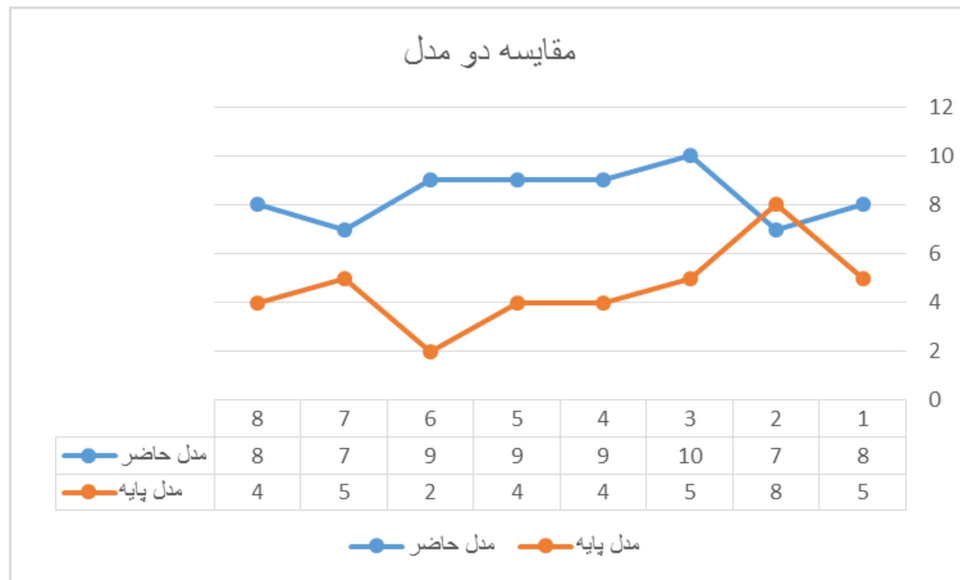


شکل ۱. گراف زمان‌دار

تعداد ۱۰ پرس‌وجو از هر پژوهش به‌طور مجموع ۲۰ پرس‌وجو به‌صورت اتفاقی انتخاب گردید و به خبرگان مشخص‌شده هر پژوهش توسط جامعه آماری سوم نمره صفر یا یک داده شد. مجموع نمرات برای هرکدام در جدول (۳) آمده است؛ که نشان می‌دهد دخالت‌دادن عامل زمان و استفاده از گراف زمان‌دار ازنظر نفر اول به میزان ۳ نمره و ازنظر نفر دوم نیز به‌اندازه ۳ نمره و ... از مدل پایه پیشی گرفته است.

جدول ۳. نتایج قضاوت ربط برای خبرگان بازیابی‌شده در هر دو مدل و مقایسه عملکرد آنها

قضاوت ربط برای خبرگان بازیابی‌شده در پژوهش حاضر	تعداد	قضاوت ربط برای خبرگان بازیابی‌شده در مدل پایه	تعداد
نفر اول	۷	نفر اول	۴
نفر دوم	۸	نفر دوم	۵
نفر سوم	۱۰	نفر سوم	۲
نفر چهارم	۹	نفر چهارم	۴
نفر پنجم	۹	نفر پنجم	۴
نفر ششم	۹	نفر ششم	۵
نفر هفتم	۷	نفر هفتم	۸
نفر هشتم	۸	نفر هشتم	۵
مجموع	۶۷		۴۶



	8	7	6	5	4	3	2	1
مدل حاضر	8	7	9	9	9	10	7	8
مدل پایه	4	5	2	4	4	5	8	5

نمودار ۱. مقایسه قضاوت خبرگان بازیابی‌شده در هر دو مدل

بحث و نتیجه‌گیری

با توجه به نوآوری به‌کاررفته در پژوهش حاضر که کاربرد مدل بازیابی اطلاعات تحلیل معنای نهان بود که درنهایت جهت بازیابی نویسندگان خبره به کار رفت، ازنظر میزان به‌دست‌آمده از هر یک سنجه‌های بازیابی اطلاعات یعنی میزان دقت در سطح پنج نتیجه اول[1]، میانگین متوسط دقت (MAP) و میانگین معکوس رتبه (MRR) به‌ترتیب با مقدار ۰/۸۹۵، ۰/۸۳۹ و ۰/۹۰۹، مدل بازیابی تحلیل معنای نهان عملکرد بهتری نسبت به مدل پایه داشت؛ و این امر به‌دلیل بهتربودن عملکرد بازیابی به‌روش کاهش ابعاد نسبت به تطابق کلیدواژه‌ای است. چون در این روش از نمایه‌سازی معنای نهان استفاده می‌شود که نوعی نمایه‌سازی مفهومی است و از روش آماری حداقل مربعات بهره می‌برد و نمایه‌سازی ذکرشده با به‌کارگیری این روش آماری استخراج می‌شود. همان طور که می‌دانیم روش‌های بسیاری برای بیان یک کلمه وجود دارد (مترادف‌ها)، بنابراین امکان دارد که کلمات پرس‌وجو با کلمات سند منطبق نشوند. همچنین اغلب کلمات معانی چندگانه‌ای دارند (چند مترادف‌ها)، پس بازیابی اطلاعات بر اساس مفهوم و معنای یک سند رویکرد بهتری است. در ال‌اس‌ای فرض این است که تعدادی ساختار نهان در کاربرد کلمات وجود دارد که تا حدی با انتخاب متنوع کلمات سد می‌شود. اس‌وی‌دی[2] برای حدس این ساختار به کار می‌رود. بردارهایی که به‌طور آماری به دست می‌آیند بیش از کلمات مجزا نشانگرهای معنا را تقویت

1. percision @ 5 (P@5)
2. SVD

می‌کنند. نتایج پژوهش‌های دیگر نیز حاکی از این مطلب است که بازیابی اسناد به‌روش تطبیق کلیدواژه‌ای پرس‌وجو با اسناد به نسبت روش ضعیف‌تری است؛ و نیز مدل بازیابی ال‌اس‌ای عملکرد بهتری در بازیابی اسناد در مجموعه بزرگ اسناد نسبت به مجموعه کوچک دارد (Magerman et al., 2010)؛ البته این موضوع در پژوهش حاضر موردبررسی قرار نگرفت. نتایج نشان دادند که تحلیل معنای پنهان عملکرد بهتری را داراست (M. W. Berry et al., 1995; M.W. Berry et al., 1995; Evangelopoulos, 2013).

یکی از محدودیت‌های پژوهش حاضر، بررسی‌نکردن متن کامل مقالات اسناد مجموعه آزمون بود. این موضوع وجود دارد که شاید حضور متن کامل مقالات بر عملکرد مدل‌های بازیابی اطلاعات تأثیر مثبتی بگذارد. لیکن نتایج پژوهش بوگرس، کوکس، و وان دن بوش (۲۰۰۸) نشان داد که عملکرد مدل‌های بازیابی اطلاعات، در شرایطی که اسناد نمایه‌شده فقط حاوی چکیده مقالات باشند، بهتر از زمانی است که اسناد نمایه‌شده حاوی متن کامل باشند(Bogers et al., 2008) . این امر مشکل مربوط به تأثیر منفی استفاده نکردن از متن کامل مقالات را تا حدود زیادی کاهش می‌دهد. نبود اصطلاح‌نامه در بطن نظام بازیابی اسناد به‌دلیل ماهیت مدل بازیابی تحلیل معنای پنهان که پیشتر توضیح داده شد از محدودیت‌های پژوهش نیست.

در مرحله بازیابی اسناد، قضاوت ربط اسناد بر مبنای عنوان، چکیده و کلیدواژه مقالات صورت می‌گرفت، اقلامی که در اغلب پژوهش‌های بازیابی اطلاعات معمول هستند. ازآنجایی‌که پژوهش حاضر قصد داشت از اسناد بازیابی‌شده در مرحله بعدی برای یافتن نویسندگان خبره بهره‌برداری کند، ارائه اسامی نویسندگان مقالات ممکن بود به سوگیری قضاوت‌کنندگان منجر شود. به همین خاطر از گنجاندن اسامی نویسندگان در فایل‌های قضاوت ربط خودداری شد. لیکن پژوهش‌های بازیابی اطلاعات می‌توانند تأثیر حضور این اقلام بر قضاوت ربط کاربران را موردتوجه قرار دهند.

با توجه به نوآوری بعدی پژوهش حاضر که دخالت‌دادن عامل زمان در خبره‌یابی بوده است که با به‌کارگیری شاخص‌های شبکه اجتماعی و قضاوت ربط نهایی انجام شده، نتایج نشان دادند که عملکرد این روش به‌طور قابل‌توجهی بهتر از مدل پایه بوده است. عامل زمان در بازیابی خبرگان دخالت داده شد تا اینکه افرادی که دیگر در قید حیات نیستند یا زمان زیادی از آخرین انتشارات آنها درزمینهٔ معین می‌گذرد بازیابی نشوند. با توجه به عمر مفید انتشارات در حوزه علم اطلاعات و دانش‌شناسی بازه زمانی ده‌ساله دخالت داده شد. پس از به‌کارگیری زمان انتشار به‌عنوان عامل تعیین‌کننده بازیابی خبرگان، کسانی که بیشترین کار مرتبط را منتشر کرده بودند به‌عنوان عامل تعیین‌کننده بعدی در نظر گرفته شدند؛ و سپس شاخص‌های خرد شبکه اجتماعی ازجمله مرکزیت درجه‌ایی، مرکزیت بینابینی، نزدیکی و نیز بردار ویژه عوامل تعیین‌کننده بعدی دیگری هستند که در پژوهش‌های علم‌سنجی به‌گونه‌ای فراوان و به‌تازگی در پژوهش‌های بازیابی خبرگان از آنها استفاده شد. ده پرس‌وجو طرح‌شده در پژوهش حاضر و نیز در پژوهش عطاپور که به‌طور تصادفی انتخاب شده بود، برای ۸ نفر که جامعه آماری دوم پژوهش

مشخص کرده بودند فرستاده شد و نتایج حاکی از آن بود که عملکرد گراف زمان‌دار و روش‌های بازیابی خبرگان با استفاده از عامل بیشترین کار مرتبط منتشر شده و عامل شاخص‌های خرد شبکه اجتماعی عملکرد بهتری داشتند.

استفاده از مدل بازیابی تخصیص دیریکله پنهان[1] در مجموعه آزمون حاضر به‌جای تحلیل معنای نهان و مقایسه آنها. مدل‌سازی موضوع از ال‌اس‌ای شروع شد، ورودی، یک ماتریس اصطلاح‌ـسند است که تعداد دفعاتی که یک اصطلاح خاص در یک سند معین می‌شود را ثبت می‌کند. در اوایل سال ۲۰۰۰، رویکرد دیگری که تخصیص دیریکله پنهان نامیده می‌شود توسعه یافت. ایده اصلی آن است که اسناد به‌عنوان ماتریس‌های تصادفی از موضوعات نهان ارائه شوند و هر موضوع توسط توزیع اصطلاحات مشخص شود.

پیمایش مجدد نظر قضاوت‌کنندگان با ارائه اطلاعات زمینه‌ای در مورد نویسندگان بازیابی‌شده و مقایسه نتایج آن با نتایج روش به‌کاررفته در این پژوهش. بدین ترتیب می‌توان در مورد پایایی قضاوت‌های انجام‌شده در پژوهش به یک بینش عمیق‌تر دست یافت.

References

Askari, A., Verberne, S., & Pasi, G. (2022). Expert Finding in Legal Community Question Answering. *Advances in Information Retrieval*, 22-30. https://doi.org/https://doi.org/10.48550/arXiv.2201.07667

Baeza-Yates, R., & Ribeiro-Neto, B. (2012). *Modern Information Retrieval*. Addison Wesley Longman Publishing Co. Inc. https://web.cs.ucla.edu/~miodrag/cs259-security/baeza-yates99modern.pdf

Balog, K. (2008). People search in the enterprise. *SIGIR Forum*, *42*(2), 103. https://doi.org/10.1145/1480506.1480526

Berry, M. W., Dumais, S. T., & O'Brien, G. W. (1995). Using linear algebra for intelligent information retrieval. *SIAM review*, *37*(4), 573-595. https://doi.org/10.1137/1037127

Berry, M. W., Dumais, S. T., & Shippy, A. T. (1995). *A case study of latent semantic indexing. Tech. Rep CS-95-271, University of Tennessee, Knoxville, January* University of Tennessee.

Bogers, T., Kox, K., & van den Bosch, A. (2008). Using citation analysis for finding experts in workgroups. Proc. DIR,

Campbell, C. S., Maglio, P. P., Cozzi, A., & Dom, B. (2003). Expertise identification using email communications. Proceedings of the twelfth international conference on Information and knowledge management,

Chaomei, C., & Paul, R. J. (2001). Visualizing a knowledge domain's intellectual structure. *Computer*, *34*(3), 65-71. https://doi.org/10.1109/2.910895

Cheng, B. (2003). *Towards Understanding Latent Semantic Indexing*.

Daud, A., Li, J., Zhou, L., & Muhammad, F. (2010). Temporal expert finding through generalized time topic modeling. *Knowledge-Based Systems*, *23*(6), 615-625. https://doi.org/10.1016/j.knosys.2010.04.008

1. Latent Drichilet Allocation (LDA)

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American society for information science*, *41*(6), 391-407. https://doi.org/10.1002/(SICI)1097-4571(199009)41:6%3C391::AID-ASI1%3E3.0.CO;2-9

Ehrlich, K., Lin, C.-Y., & Griffiths-Fisher, V. (2007). Searching for experts in the enterprise: combining text and social network analysis. Proceedings of the 2007 international ACM conference on Supporting group work,

Evangelopoulos, N. E. (2013). Latent semantic analysis. *Wiley Interdisciplinary Reviews: Cognitive Science*, *4*(6), 683-692. https://doi.org/10.1002/wcs.1254

Fang, H., & Zhai, C. (2007). Probabilistic models for expert finding. Advances in Information Retrieval: 29th European Conference on IR Research, ECIR 2007, Rome, Italy, April 2-5, 2007. Proceedings 29,

Fu, Y., Xiang, R., Zhang, M., Liu, Y., & Ma, S. (2006). A PDD-Based searching approach for expert finding in intranet information management. Information Retrieval Technology: Third Asia Information Retrieval Symposium, AIRS 2006, Singapore, October 16-18, 2006. Proceedings 3,

Kanhabua, N., & Nørvåg, K. (2010). Determining Time of Queries for Re-ranking Search Results. European Conference on Research and Advanced Technology for Digital Libraries,

Li, J., Tang, J., Zhang, J., Luo, Q., Liu, Y., & Hong, M. (2007). Eos: expertise oriented search using social networks. Proceedings of the 16th international conference on World Wide Web,

Ligtenberg, W., & Pei, Y. (2017). Introduction to a temporal graph benchmark. *arXiv preprint arXiv:1703.02852*. https://doi.org/10.48550/arXiv.1703.02852

Macdonald, C. (2009). *The voting model for people search* University of Glasgow. http://theses.gla.ac.uk/id/eprint/609

Mageman, T., Van Looy, B., & Song, X. (2010). Exploring the feasibility and accuracy of Latent Semantic Analysis based text mining techniques to detect similarity between patent documents and scientific publications. *Scientometrics*, *82*(2), 289-306. https://doi.org/10.1007/s11192-009-0046-6

Mathews, L. K., & Kanmani, S. D. (2012). A survey on temporal information retrieval systems. *International Journal of Computer Applications*, *58*(4).

Michail, O. (2015). An introduction to temporal graphs: An algorithmic perspective. In *Algorithms, Probability, Networks, and Games: Scientific Papers and Essays Dedicated to Paul G. Spirakis on the Occasion of His 60th Birthday* (pp. 308-343). Springer.

Omidvar, A., Garakani, M., & Safarpour, H. R. (2014). Context based user ranking in forums for expert finding using WordNet dictionary and social network analysis. *Information Technology and Management*, *15*, 51-63. https://doi.org/10.1007/s10799-013-0173-x

Salton, G., Wong, A., & Yang, C.-S. (1975). A vector space model for automatic indexing. *Communications of the ACM*, *18*(11), 613-620. https://doi.org/10.1145/361219.361220

Sanderson, M. (2010). Test Collection Based Evaluation of Information Retrieval Systems. *Foundations and Trends® in Information Retrieval*, *4*(4), 247-375. https://doi.org/10.1561/1500000009

Sarmad, Z., Bazargan, A., & Hejazi, E. (2023). *Research methods in behavioral sciences* (Vol. 1). Agah Publication. https://www.agahbookshop.com/%D8%B1%D9%88%D8%B4%E2%80%8C%D9%8

7%D8%A7%D9%8A-%D8%AA%D8%AD%D9%82%D9%8A%D9%82-%D8%AF%D8%B1-%D8%B9%D9%84%D9%88%D9%85-%D8%B1%D9%81%D8%AA%D8%A7%D8%B1%D9%8A_202

Schwartz, M. F., & Wood, D. C. (1993). Discovering shared interests using graph analysis. *Communications of the ACM*, *36*(8), 78-89. https://dl.acm.org/doi/pdf/10.1145/163381.163402

Serdyukov, P., & Hiemstra, D. (2008). Modeling documents as mixtures of persons for expert finding. Advances in Information Retrieval: 30th European Conference on IR Research, ECIR 2008, Glasgow, UK, March 30-April 3, 2008. Proceedings 30,

Zhang, J., Ackerman, M. S., & Adamic, L. (2007). Expertise networks in online communities: structure and algorithms. Proceedings of the 16th international conference on World Wide Web,

Zhang, J., Tang, J., & Li, J. (2007). Expert finding in a social network. Advances in Databases: Concepts, Systems and Applications: 12th International Conference on Database Systems for Advanced Applications, DASFAA 2007, Bangkok, Thailand, April 9-12, 2007. Proceedings 12,

Zhang, J., Tang, J., Liu, L., & Li, J. (2008). A Mixture Model for Expert Finding. Advances in Knowledge Discovery and Data Mining, Berlin, Heidelberg.

Zhou, D., Orshanskiy, S. A., Zha, H., & Giles, C. L. (2007). Co-ranking authors and documents in a heterogeneous network. Seventh IEEE international conference on data mining (ICDM 2007),